A blueprint for symptom scales and responses: measurement and reporting

K W Wyrwich, V M Staebler Tardino

Gut 2004;53(Suppl IV):iv45–iv48. doi: 10.1136/gut.2003.034348

Measurement of symptom status among patients with GORD has, to date, been conducted in a variety of ways. We applied a cognitive psychology framework, as well as reviewed the current psychometric literature, to develop a blueprint for the best methods of measuring and reporting symptoms of patients with gastro-oesophageal reflux disease (GORD). Our review suggests that the seven point scale with word anchors for each point is likely to afford better reliability, improved sensitivity to change in GORD symptoms, and greater ease in administration compared with visual analogue or other types of response scales.

SUMMARY

Four cognitive steps have been outlined that respondents must traverse when answering questionnaire items. These are: (1) interpret what is being asked; (2) retrieve relevant information; (3) make a summary judgment; and, finally, (4) convey that judgment to the examiner. We applied this cognitive psychology framework, as well as reviewed the current psychometric literature, to develop a blueprint for the best methods to measure and report symptoms of patients with gastro-oesophageal reflux disease (GORD). In order to create symptom scales that simplify these cognitive tasks for GORD patients, research is needed into wording and response options that patients can easily understand that evaluate the main symptoms of GORD. Providing GORD patients with simple memory aids, such as anchoring dates when assessing symptoms over a relevant time span, can ease the retrieval process of pertinent information. Previous work detailing how 7 ± 2 choices are the usual limits of our human capacity to process information offers guidance for the number of symptom response options that can best convey summary judgments. Our review suggests that the seven point scale with word anchors for each point is likely to afford better reliability, improved sensitivity to change in GORD symptoms, and greater ease in administration compared with visual analogue or other types of response scales. When reporting GORD treatment outcomes, data on the proportion of patients achieving a threshold for clinically important change in their symptoms supplies relevant information that cannot be gleaned from group mean change assessments.

INTRODUCTION

Measurement of symptom status among patients with GORD has, to date, been conducted in a variety of ways. In 1998, Moyer and Fendrick reviewed and described the many methods used to assess health related quality of life (HRQoL).¹ The items in the reviewed questionnaires sampled from a variety of domains and used an assortment of formats and response options, with no consistent configuration in their presentation. It is interesting to note that most of the questionnaires used over the past decade were developed by physicians dedicated to their quest to better quantify, and subsequently reduce, disease related suffering of their patients (Sharma and colleagues² in this supplement (*see page iv58–iv65*)).

Yet at a time when reliable and valid measures are needed to evaluate the change in GORD patients seeking reduction of their symptoms through known and emerging treatments, uncertainty remains on the best methods for measuring symptom status among GORD patients. In addition, the need to understand the worldwide burden of this disease requires both a global understanding of the effects of reflux disease and a standardised manner of reliably and validly measuring and comparing results within and across cultures, countries and geographic regions.

This report provides a blueprint for the development of questionnaire items to measure a hierarchy of GORD treatment outcomes, such as symptom status and HRQoL, by assimilating the results from others who have investigated the most appropriate measurement methodologies. This report also presents a blueprint for the most suitable methods to evaluate and report relevant changes in status over time in these outcome measures as a result of treatment. These evaluation and reporting methods should provide the many stakeholders in the treatment of GORD with a useful process for interpreting treatment effects.

MEASUREMENT

In his often cited review, Roger Tourangeau³ outlined four cognitive steps that people must traverse when responding to questionnaire items. These are: (1) interpret what is being asked; (2) retrieve relevant information; (3) make a summary judgment; and, finally, (4) convey that judgment to the examiner. These

Abbreviations: GORD, gastro-oesophageal reflux disease; HRQoL, health related quality of life; GSRS, gastrointestinal symptoms rating scale

See end of article for authors' affiliations

Correspondence to: Dr K W Wyrwich, Saint Louis University, 221 North Grand Ave, Saint Louis, MO 63103, USA; wyrwichk@slu.edu

four stages in patient response solicitation provide insights for the development of GORD questionnaire items that ease respondent burden (for example, decrease the cognitive complexity of the task) and, in doing so, improve the reliability and completeness of patient responses.

Step 1: Interpret what is being asked

In order to create appropriate scales that simplify the patient's task of interpreting what is being asked, qualitative research is needed into the important GORD symptoms and the HRQoL aspects affected. In addition, the most appropriate response options among patients with GORD must also be qualitatively investigated.⁴ Mechanisms for accomplishing this qualitative inquiry are concurrent or retrospective approaches, such as cognitive interviewing or think aloud protocol interviewing techniques.⁵ Clinicians who treat GORD patients often have an exceptional and comprehensive understanding of their symptoms and the important HRQoL aspects of the disease. However, exploration of these concepts among GORD patients is necessary in order to make the response options most meaningful and user friendly to the respondents. Moreover, patients are the best informants on response options that meaningfully identify the frequency and intensity of their disease,⁴ but this qualitative research is noticeably lacking in reflux disease.

Cultural adaptations

As constructs may be interpreted differently and vary in relevance for particular cultures, it is also important that conceptual differences in symptom assessment across cultures—even those cultures with the same primary language—be qualitatively reassessed when existing instruments are adapted.⁴ This includes qualitatively researching new wordings and cultural adaptations among users that are most meaningful to a particular culture and then assessing the psychometric equivalence of new adaptations with other versions of the questionnaire. The rationale for conducting cultural adaptations in this manner is illustrated by the report of Talley *et al* on developing a new dyspepsia impact scale, the Nepean dyspepsia index.⁶ In the process of assembling important HRQoL aspects of this symptom, the international collaboration of specialists identified the five most important lifestyle limitations caused by dyspepsia from each of their respective countries. While Italians rated their loss of enjoyment of meals with family and friends as the most important limitation, this disease limitation was not even listed by Australians. Hence cultural differences between countries may require different items to optimally capture the impact of symptoms and HRQoL relevant to any particular cultural group.

Similarly, the options from which patients select when responding to symptom and HRQoL questions also need to be adequately investigated for new cultural adaptations. The extensive work by the World Health Organisation Quality of Life (WHOQOL)⁷ investigators demonstrates that although different cultures may use the same primary language, differences in the meaning and ranking of word descriptors may produce psychometrically non-comparable results if this important qualitative phase is ignored.

Step 2: Retrieve relevant information

Tourangeau's second cognitive step explores patients' abilities to remember important GORD symptoms and HRQoL issues to recover pertinent frequency and intensity measurements.³ Unfortunately, patients generally are not motivated to complete daily diaries where symptom data can be collected systematically and consistently. Instead, diaries tend to be either ignored or completed through a process called "hoarding" where a series of entries are completed at one time.⁸ Hence symptom and HRQoL questionnaires

typically rely on patients' memories of their condition, and not written records.

In a recent report on evaluation questionnaires, Schwartz and Oyserman presented a comprehensive review of key lessons in autobiographical memory research.⁹ Current theory on autobiographical memory proposes that our memories are organised into hierarchical networks¹⁰ that group extended periods of time into meaningful life events (for example, when I was a child, when I was in college, when I was married to Joanne, etc). Within each of these groups are additional substructures that further subdivide. For example, the college years might be organised into undergraduate and graduate school, if these are meaningful subdivisions to the individual. These subdivisions and further partitionings continue into smaller and smaller components with specific events (for example, severe heartburn symptoms during my son's wedding) being at the lowest level of this hierarchy. Memory of these specific events however requires that they generally be rather unusual (for example, son's wedding).¹¹ Otherwise, specific events merely blend into global knowledge-like representations at a higher level of the memory network (for example, I had heartburn often when I worked for the telephone company).

Applying this theory on autobiographical memory structures, it is crucial that GORD symptom and HRQoL questionnaire items supply patients with useful ways to recover these relevant groups and events. As it is difficult to recall distant events that are not unusual, shorter recall periods improve the accuracy of recalled symptoms and HRQoL effects. Furthermore, providing appropriate recall cues, such as dates and cues pertaining to unusual events (heartburn so bad that you had stay home) can help patients in their accurate recall of GORD related events.

Steps 3 and 4: Make a summary judgment and convey that judgment to the examiner

After retrieving relevant events, GORD patients must then convert and report these events in the appropriate formats. Most commonly, these formats are either open ended visual analogue scales (VAS) or modified Likert scales. Open ended response formats allow patients to report the frequency or intensity of their symptoms and HRQoL effects in their own words. Although this can impart rich qualitative data on each patient's health status, there are important limitations. Firstly, this format tends to yield more missing data, compared with modified Likert scales, because patients: (a) do not like to create their own responses; and (b) some patients cannot write or are not comfortable with their writing skills. Secondly, an open ended response takes more cognitive and physical effort to complete and increases task complexity. A third limitation to the format is the non-comparability of patient responses when data are compiled. That is, when asked: "How much have you been bothered by heartburn during the past week?", one patient's report of "every day of the week" cannot easily be compared with another patient's response of "a lot."

Visual analogue scales

VAS ask patients to respond to GORD symptom questions by placing a mark on a straight line that is typically 10 cm long. The line is often anchored with "0" at one end and "100" at the other. Therefore, when a patient responds on this scale to an item like: "On a scale from 0 to 100, how much have you been bothered by heartburn during the past week?" by placing a mark 7.8 cm above the "0," their response can be converted into the number 78 on this 0–100 scale. Hence the VAS format yields theoretically continuous response data which simplifies computations and analyses of these data, and often produces greater variance in the data compared with other response formats.¹²

Despite these benefits, there are several drawbacks to VAS. Primary among these limitations are the documented and repeated problems that patients demonstrate in both learning and understanding how a mark on a straight line can reflect a construct like the intensity of their pain.¹³ These learning and comprehension difficulties are especially great among older adults and those with less education.^{13 14} In addition, problems result from the lack of uniform meaning among patients about any particular point on the line. That is, can the meaning of the 7.8 cm mark be compared between a GORD patient who suffers from severe heartburn nearly every day and another GORD patient with only occasionally mild episodes?

Modified Likert scales

Modified Likert scales have been a popular option for survey responses since the introduction of the Likert scale in 1932.¹⁵ Originally offering only five response options (strongly disapprove, disapprove, uncertain, approve, and strongly approve), these response scales have been adapted and expanded to include at least 19 response options per item when measuring latent constructs.¹⁶ The discrete number of response options, coupled with word descriptors for each response choice, yields more complete data than either open ended or VAS options.¹⁷

The optimal number of response options for a modified Likert scale remains a topic for debate. However, several key studies offer evidence of the superiority of the seven point scales. First among these is George A Miller's (1956) pivotal work, "The magic number seven plus or minus two: some limits of our capacity for processing information".¹⁸ Using several examples from the psychophysics literature, Miller presented evidence that there are often far too many incoming stimuli to process and store in our memories. Therefore, we must organise the information by "chunking" it together and, in so doing, making better use of our cognitive resources. He went on to demonstrate that short term memory can hold only five to nine chunks of information (seven plus or minus two), where a chunk is any meaningful unit. A chunk can refer to digits, words, chess positions, or people's faces, and also offers guidance on the optimal number of modified Likert scale response options.

Additional studies allow us to further examine the range of five to nine response options. In a series of papers exploring the results from their systematic collection of responses to the Florida Scale of Civic Beliefs using 2–19 point modified Likert scales, Mattell and Jacoby demonstrated that: (a) six or seven point scales appear to be the optimal length for information recovery and reproducibility (test–retest reliability); and (b) considerably fewer respondents (7% *v* 20%) used the middle response option on 7–19 point scales compared with three and five point response scales.¹⁶ Hence "fence sitters" who are uncomfortable with the dichotomy introduced by an even number of response options have a happy home on a seven point scale, although the Mattell and Jacoby results show that they are less likely to make this middle choice compared with a five point response scale. Walter and colleagues¹⁹ conducted bootstrapping, *t* tests, and multiple regression analyses using HRQoL data, and concluded that if a measure uses a discrete response scale, it "should be treated as continuous if it has seven or more categories and as ordinal otherwise (p 156)". Finally, Guyatt and colleagues compared VAS and seven point scales for measuring change in respiratory rehabilitation patients and concluded that both provided a sensitive process for measuring patient change, with no statistically significant differences between the two scales. However, due to the ease of administration and

interpretation, these researchers recommended the seven point scale over the VAS.¹⁷

With any modified Likert scale, it is important that the label or word descriptors used for each discrete response point provide a meaningful hierarchy for the patients who are responding.⁴ As previously stated, these response options should be solicited through qualitative research methods (for example, cognitive pretests). In addition, item response theory and the resulting item characteristic curves provide further avenues to confirm the hierarchical and linear scaling for any modified Likert response set. Finally, special care should be exercised when selecting the response option descriptors so that the first and last descriptors fully capitalise on the benefits of a seven point modified Likert scale, and provide sufficient psychological width to match the experiences of respondents.¹⁵

REPORTING RESULTS

Evaluating and reporting GORD treatment effects from the longitudinal collection of symptoms data poses a methodological quagmire. Analysis of variance (ANOVA) and *t* tests have a longstanding tradition as the statistical tests that ascertain significant changes. These statistical techniques, with their focus on *p* values, assure us that the observed statistically significant changes are likely to not be due to chance. The results however are dependent on the sample size and variation. Therefore, these evaluations of change can only determine if there is a likely chance that the symptom scale group means are different across time but provide no indication of the relevance, if any, of these differences to the patients. Furthermore, most GORD treatment outcomes are focused towards change in individual participants yet evaluation of intraindividual change is lost when only group means are compared.²⁰

Interpreting results from group mean changes

We will use the recent publication, "Quality of life in patients with heartburn but without esophagitis: effects of treatment with omeprazole," to illustrate this evaluation dilemma.²¹ This study reported a randomised controlled trial comparing the impact of treatment on HRQoL for patients receiving 10 and 20 mg of omeprazole, and placebo. The gastrointestinal symptoms rating scale (GSRS) was used for this evaluation, as well as other measures, at baseline, after pre-entry endoscopy, and again after four weeks of treatment.²² In their results comparing changes from baseline, where randomisation had resulted in nearly equivalent group mean baseline values, compared with the four week follow up assessment, the placebo group (*n* = 80) had a mean reduction of 0.55 points on the 1–14 point scale of the GSRS reflux dimension while the two omeprazole groups had mean reductions of 1.02 and 1.27 points, respectively. Compared with the placebo group, these changes were statistically significant at the 0.0001 and 0.003 levels, respectively. However, in the constipation dimension of the GSRS where, again, similar baseline mean values were present across the three groups, mean reductions were 0.01 in the placebo group, and 0.21 and 0.13 in the respective 10 and 20 mg omeprazole groups. These changes were not statistically significant (*p* > 0.05).

Although researchers have come to acknowledge that these results imply that the treatment was effective in significantly reducing reflux (heartburn and acid regurgitation) but not constipation, they should question several issues related to statistical significance testing, such as the following.

- (1) If only a few more patients had been enrolled, would the mean constipation dimension change difference between omeprazole and placebo have reached significance?

- (2) How small could the pre-post reflux dimensional change be and still achieve statistical significance?
- (3) Was the statistically significant change in the reflux dimension meaningful or important to the enrollees?

Indeed, when we come to appreciate that with the appropriately powered study, mean change differences as small as those observed in the GSRS constipation dimension could reach significance, we also realise that statistically significant results do not tell us much about meaningful symptom change among the patients enrolled in the study. This is because statistically significant group changes do not necessarily imply meaningful or interpretable differences for individual patients.²³

Interpreting results for the proportion of patients improving

As the example above illustrates, symptom and HRQoL measurement requires a different methodology of reporting and evaluating treatment changes. A more appropriate and useful strategy is to state the proportion of patients achieving a threshold for clinically important change in their symptoms. Such a report can also be evaluated for statistically significant comparisons of the percentage reaching the threshold for improvement to evaluate treatment efficacy. But more importantly, the percentage reaching the clinically important difference threshold supplies clinicians, patients, and other stakeholders with relevant information that cannot be gleaned from group mean change assessments.²⁴ Having a standard for assessing clinically important change presents a meaningful manner to classify a patient's status as improved, stable, or worsened. It also subsequently improves estimation methods for the likelihood of important symptom change through event modelling procedures, such as proportional hazards regression, polytomous regression, and logistic regression.

How can such a threshold for symptoms and HRQoL measures be established for the treatment of GORD? Both anchor and distribution based procedures²⁵ for determining clinically important difference standards among patients and clinicians have been developed, and refinement of these procedures continues as researchers and practitioners seek to establish these important thresholds across many disease specific and generic health status measures. Anchor based approaches examine patient score changes that correspond with other acknowledged measures of clinically important improvements (the anchors). Distribution based methods, such as effect size and the standard error of measurement, use statistical parameters to gauge important levels of change. Likewise, treatment outcomes can be assessed using an absolute threshold, such as the absence of all heartburn for an entire week, regardless of the patient's initial symptom status. Despite the varied methods for establishing a threshold as the standard for determining a clinically important difference by which the patient's progress is measured and reported, agreement and use of the standard is both sensible and essential. In doing so, use of a clinically important change standard will allow symptom and HRQoL measurements to become increasingly interpretable, and a universally accepted component in assessing not only treatment efficacy but also quality of patient care.²⁶

Authors' affiliations

K W Wyrwich, Departments of Research Methodology and Health Services Research, Saint Louis University, Saint Louis, Missouri, USA
V M Staebler Tardino, Department of Psychology, Saint Louis University, Saint Louis, Missouri, USA

REFERENCES

- 1 Moyer C, Fendrick A. Measuring health-related quality of life in patients with upper gastrointestinal disease. *Dig Dis* 1998;16:315-24.
- 2 Sharma N, Donnellan C, Preston C, et al. A systematic review of symptomatic outcomes used in oesophagitis drug therapy trials. *Gut* 2004; 53(suppl IV):iv58-65.
- 3 Tourangeau R. Cognitive sciences and survey methods. In: Jabine T, Straf M, Tanur J, et al, eds. *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press, 1984:73-100.
- 4 Stewart A, Napoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. *Med Care* 2000;9(suppl II):II-102-24.
- 5 Patton MQ. *Qualitative research and evaluation methods*, 3rd edn. Thousand Oaks: Sage Publications, 2002.
- 6 Talley NJ, Haque M, Wyeth J, et al. Development of a new dyspepsia impact scale: the Nepean Dyspepsia Index. *Aliment Pharmacol Ther* 1999;13:225-33.
- 7 Szabo S, Orley J, Saxena S. An approach to response scale development for cross-cultural questionnaires. *Eur Psychol* 1997;2:270-6.
- 8 Stone AA, Shiffman S, Schwartz JE, et al. Patient non-compliance with paper diaries. *BMJ* 2002;325:1193-4.
- 9 Schwartz N, Oyserman D. Asking questions about behavior: cognition, communication, and questionnaire construction. *Am J Eval* 2001;22:127-60.
- 10 Belli R. The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory* 1998;6:383-406.
- 11 Conway M. *Autobiographical memory: An introduction*. Buckingham, UK: Open University Press, 1990.
- 12 Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: Visual analogue scales compared to rating scales in a Likert format. *Psychol Rep* 1995;77:475-80.
- 13 Ferraz M, Quaresma M, Aquino L, et al. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *J Rheumatol* 1990;17:1022-4.
- 14 Huskisson E. Measurement of pain. *Lancet* 1974;2:1127-31.
- 15 Wyatt R, Meyers L. Psychometric properties of four 5-point Likert-type response scales. *Educ Psychol Meas* 1987;47:27-35.
- 16 Mattell M, Jacoby J. Is there an optimal number of alternatives for Likert-scale items?: Effects of testing time and scale properties. *J Appl Psychol* 1972;56:506-9.
- 17 Guyatt G, Townsend M, Berman L, et al. A comparison of Likert and visual analogue scales for measuring change in function. *J Chronic Dis* 1987;40:1129-33.
- 18 Miller G. The magic number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 1956;63:81-97.
- 19 Walter S, Campbell M, Lall R. Design and the analysis of trials with quality of life as an outcome: A practical guide. *J Biopharm Stat* 2001;11:155-76.
- 20 Wyrwich K, Wolinsky F. Identifying meaningful intra-individual change standards for health-related quality of life measures. *J Eval Clin Pract* 2000;6:39-49.
- 21 Havelund T, Lind T, Wiklund I, et al. Quality of life in patients with heartburn but without esophagitis: Effects of treatment with omeprazole. *Am J Gastroenterol* 1999;94:1782-9.
- 22 Dimenas E, Glise H, Hallerback B, et al. Well-being and gastrointestinal symptoms among patients referred to endoscopy owing to suspected duodenal cancer. *Scand J Gastroenterol* 1995;30:1046-52.
- 23 Wyrwich K, Nienaber N, Tierney W, et al. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* 1999;37:469-78.
- 24 Wyrwich K, Tierney W, Wolinsky F. Using the standard error of measurement to identify important intra-individual change on the Asthma Quality of Life Questionnaire. *Qual Life Res* 2002;11:1-7.
- 25 Lydick E, Epstein R. Interpretation of quality of life changes. *Qual Life Res* 1993;2:221-6.
- 26 Clancy C, Eisenberg J. Outcomes research: measuring the end results of health care. *Science* 1998;282:245-6.